Aug 10th, 12:00 AM

What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL

Matthew Gifford
Fingerpaint, mgifford1717@gmail.com

Tuncay Bayrak
Western New England University, tbayrak@wne.edu

What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL

Emergent Research Forum (ERF)

Matt Gifford
Fingerpaint, Saratoga Springs, NY
matthewgifford17@gmail.com

Tuncay Bayrak
Western New England University
tbayrak@wne.edu

Abstract

Although analytics has become commonplace within the sports industry and is growing within the National Football League, there is limited evidence available regarding the construction of a prediction model to determine which factors and more specifically which team statistics have the largest impact on winning. In this study, we strive to generate multiple statistical models to quantify influence of team statistics on regular season wins, evaluate the created models based on accuracy to determine the best predictive model, and validate the created model by applying the regression to the 2018 NFL regular season games and compare to the actual season standings.

Keywords

Sport analytics, national football league, statistical models.

Introduction

In Week 11 of the 2012 NFL season, the Atlanta Falcons were trailing 19-16 against the Arizona Cardinals, as the offense took the field with 9:44 remaining in the 4th quarter. After Matt Ryan recorded four interceptions and Jason Snelling recorded a fumble through the first three quarters of play, the Falcons were currently at a five to one deficit in turnovers. Starting on their own 30 yard line, the Atlanta offense would manage to string together a 3 minute, 70 yard drive capped off by a rushing touchdown by Michael Turner in order to take the lead. Matt Ryan would later commit his fifth interception but regardless the Falcons would come away with a 23-19 win. How is it possible that the Falcons won the game with a six to one difference in turnovers? Is this game an outlier or do turnovers not have a significant impact on winning a game? Would an additional turnover have pushed the advantage to the Cardinals? We strive to answer these questions by developing a predictive model determining the likelihood of a win.

Every NFL player, coach and fan has their opinion on which team statistic or football metric provides the greatest insight to the level of success in a game or season. As analytics continues to grow in the NFL, the majority of current evidence remains anecdotal, leaving much to still be discovered and learned. Through the creation of a predictive model, we can determine with measurable accuracy how different team statistics interact with the outcome of a game. Which team statistics are most important – offensive stats such as passing yards, defensive stats such as turnovers forced or other factors such as how many wins a team already has going into a game? Models such as the ones created in this study can change the way teams play the game and how they prepare.

In this paper, we strive to accomplish the following objectives:

1) Collect and compile a large sample of recent NFL regular season game team statistics,
2) Generate multiple statistical models to quantify influence of these team statistics on regular season wins,
3) Evaluate the created models based on accuracy to determine the best predictive model, and
4) Validate the created model by applying the regression to the 2018 NFL regular season games and compare to the actual season standings.

Background

Although analytics has become commonplace within the sports industry and is growing within the National Football League, there is limited evidence available regarding the construction of a prediction model to determine which factors and more specifically which team statistics have the largest impact on winning.

As displayed by Leigh Steinberg, the origin of analytics in sports can be traced to Oakland Athletics' General Manager, Billy Beane who used statistics and the analysis of data to find value players when constructing the 2002 Athletics roster. His philosophy remained that a team compiled of players with high on-base percentages were more likely to score runs thus translating into more wins (Steinberg, 2015). Analytics has significantly grown since then and is now making headway in professional NFL Football. The Philadelphia Eagles currently have an analytics tandem which are used as resources over the game-day communication for critical football decisions. Through the use of data which they have at their disposal, the team can make more educated decisions on when to attempt for two-point conversion after a touchdown or keep the offense on the field on fourth down (McManus, 2017).

ESPN Staff Writer, Bill Barnwell set out to determine which team statistics better evaluate the quality of a team and the probability of future wins aside from their current season record. He believed one of the best metrics was Defense-adjusted Value Over Average (DVOA), a creation of Aaron Schatz from Football Outsider. DVOA compares the success of a team during a given play with the expected result when factoring the situation, down, distance and the opponent. He also noted the importance of the point differential between the team and opponents faced during the season as well as the team's record in close games (Barnwell, 2017). With the use of Minitab, Kevin Rudy took a different approach, focusing on turnovers and how the season as a whole is impacted. The results showed that 44% of the variation in a team's winning percentage was explained by the means of a team's turnover differential (Rudy, 2014). In order to determine the importance of passing and rushing in comparison to team performance, Dr. Ed Feng developed scatterplots based on efficiency during the regular season. Efficiency refers to the yards per play gained on offense subtracted by yards per played allowed on defense. Dr. Feng found determined passing efficiency to be more significant since many playoff and Super Bowl contenders excelled in passing while rushing efficiency is extremely scattered. Approximately 88% of playoff teams from 2003-2012 gained more yards per play than they allowed (Feng, n.d.).

Three MIT students developed a predictive, binary logistic regression with data from the 2000-2011 NFL seasons to understand the most influential factors of field goal success. Using a combination of SPSS and SYSTAT, the students concluded that almost all environmental factors had a significant impact on field goal success although the most important was not the kick distance but rather the altitude of the stadium. Additionally, all situational or psychological factor were insignificant including factors included whether the kick was attempted in a regular season or postseason game, a home or away game, a high pressure or low pressure situation and whether a timeout was call before the kick was attempted ("icing the kicker") (Johnson, 2013).

Lastly, a member of the JMP staff developed a neural network via the JMP software to predict wins based on a variety of variables include statistics from the previous season and the salary teams pay each position. The individual determined the top predictors were the team's number of wins and point differential from the previous season. The model also revealed that spending more money on offense rather than defense leads to improvement (JMP, 2017).

Research Method

Sample

Through the use of Pro Football Reference and its extensive historical database of football metrics and statistics, we compiled a sample of sixteen seasons worth of data to analyze and evaluate. The sample was compiled of all regular season games from the 2002-2017 seasons. We chose to begin the sample with the 2002 season as it marks the introduction of the Houston Texans as the 32nd NFL team while still being

relevant enough to the way the game is played today. With this sample, we have the same number of teams competing and games occurring in each year of data. In total, the sample includes a total of 4,096 games.

Data Preparation

Before the creation of the two models, we took many precautionary measures to ensure the data was properly cleaned. The first step was to remove ties from the data set. Throughout the 4,096 games which took place during the sixteen seasons only seven ties occurred. With this accounting for about 0.17% of the sample, we ruled these insignificant and removed them from the data set. Additionally, in order to properly create the model, we converted the target variable which was a win or loss into a binary, 1 or 0. Lastly, in order for other variables to be better recognized by SAS Enterprise Miner, dummy variables were developed for Overtime and the location of the game (Home or Away).

Model Creation

In order to determine which team statistics have the greatest impact on the outcome of a game, we selected two models to build, one classification model and one estimation/prediction model. The first model created was a decision tree (Figure 1). A decision tree is a powerful tool which can assist in determining which teams in the NFL are more likely to win based on the analyzed data. The analysis does so by determine which attributes a team should possess in order to have the best chance at winning a game. Therefore, by understanding the most important attributes to winning and which team possess them, we can conclude the outcome of particular game. Decision tree takes all of the data and runs it through individual variables, determines a cutoff value for the variable and then splits the data based off of the value.

With the completion of the decision tree, we generated a binary logistic regression as our estimation/prediction model (Figure 1). Since the target variable, winning, is binary, as it is limited to winning and losing, a logistic regression is a suitable model to run. In running this regression, we will be able to determine which variables or team statistics have the largest impact in determining the target variable of winning. In the process, this analysis will display which variables are meaningless and negligible and therefore should not be included in the model. Thus, the result is a refined model consisting of important variables which can aid us in prediction. Based on the refined model created in the analysis, we can predict the outcome of a regular season game from the team statistics. As seen from the output, the p-value of all variables listed is less than 0.05, which is our chosen alpha, therefore they will all be included in the final. As a note, the variables listed below is a subset of the original list of variables. Through the estimates column, SAS displays the coefficients for each variable within the model which are all part of the equation to determine the predictive value.

Data Analysis and Results

Through the decision tree model, we determined that the most important team statistic in determining the winner of an NFL game is offensive turnovers. The four most important variables in order include offensive turnovers (turnovers lost), defensive turnovers (turnovers forced), def rush yards (rushing yards allowed) and off total yards (total yards of offense). According to the model, only 32% of teams won a game when turning the ball over at least twice on offense. Additionally, approximately 83% of teams won if they turned the ball over on offense no more than once and forced two or more turnovers on defense.

Through the binary logistic regression, the most important variable is offensive turnovers which has a heavy negative effect (-0.9686) on the outcome while defensive turnovers is slightly less important and has a strong positive impact (0.9124). Other notable takeaways include that offensive rushing yards and defensive passing yards are included in the final model but not the opposite and home field advantage does exist (Table 1).

After comparing the Decision Tree and Binary Logistic Regression models, the Binary Logistic Regression is deemed to be the final model since it displays a lower average squared error. The average squared error for the Decision Tree is approximately 0.1693 while the Binary Logistic Regression is 0.1168 (Table 2).

Validation

Upon completion of the Binary Logistic Regression, the same fourteen metrics from the 2018 NFL Regular Season were collected. Using the coefficients developed by the model, each team within each game was assigned a Win Value. Win Values were then compared for every game within the season and the team with the higher Win Value was assigned a win with the other team being assigned a loss. After all games were completed, regular season records were determined and placed in league standings, abiding by NFL Tiebreaking Procedures. Table 3 displays the final result with the comparison of the model's predicted standings with the actual standings and results from the 2018 regular season. With two ties occurring in the 2018 season, we used 254 of the 256 games for comparison. After comparison, we determined that the result of the games were correctly predicted approximately 83.07% of the time. Additionally, 25 of the 32 teams were properly ordered within the standings.

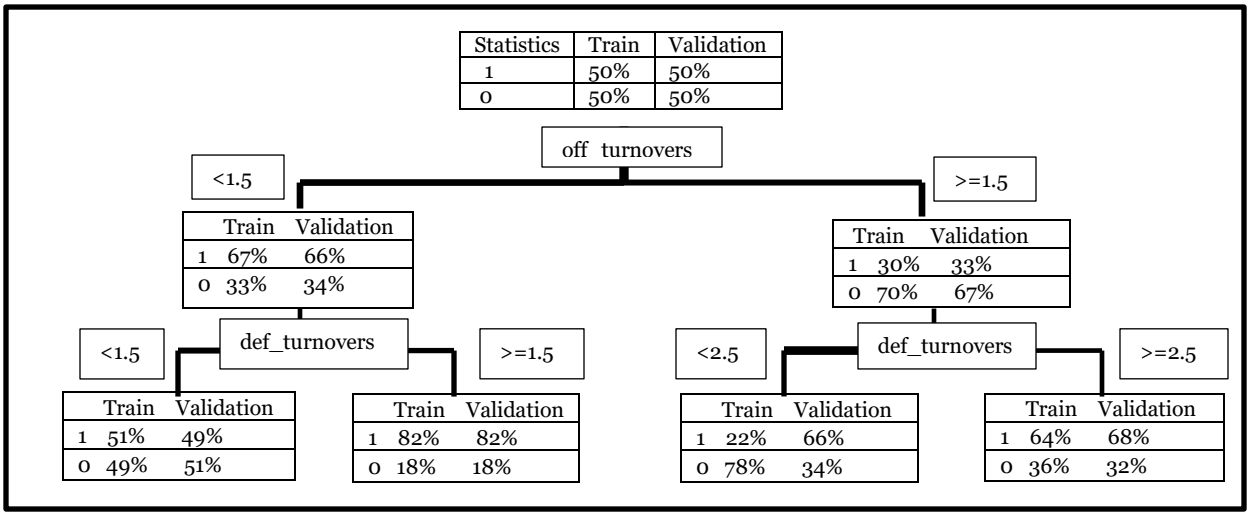

Figure 1. Decision Tree

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Std.Error	Wald ChiSq	Pr>ChiSq	Std.Estimates	Exp (Est)
Intercept	1	0.0488	0.3235	0.02	0.8802		1.050
Def_1st_down	1	-0.0428	0.0158	7.35	0.0067	-0.1186	0.958
Def_pass_yds	1	0.00651	0.00108	36.28	<.0001	0.2782	1.007
Def_total_yds	1	-0.0145	0.00131	121.83	<.0001	-0.6760	0.986
Def_turnovers	1	0.9124	0.0435	440.87	<.0001	0.6785	2.490
Home 0	1	-0.2660	0.0462	33.19	<.0001		0.766
losses	1	-0.0827	0.0165	25.12	<.0001	-0.1333	0.921
Off_1st_down	1	0.0331	0.0162	4.16	0.0415	0.0910	1.034
Off_rush_yds	1	0.00794	0.00106	55.89	<.0001	0.2291	1.008
Off_total_yds	1	0.0080	0.00099	66.13	<.0001	0.3780	1.008
Off_turnovers	1	-0.9686	0.0444	475.90	<.0001	-0.7211	0.380
Wins	1	0.0845	0.0161	27.55	<.0001	0.1387	1.088

Table 1. Logistic Regression.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Average Squared Error
Y	Reg	Reg	Binary Logistic Regression	Outcome	Outcome	0.116798
	Tree	Tree	Decision Tree	Outcome	Outcome	0.169327

Table 2. Model Comparison

	Actual			Predicted			Actual			Predicted	
AFC East	**W**	**L**	**T**	**W**	**L**	**NFC East**	**W**	**L**	**T**	**W**	**L**
New England P*	11	5	0	10	6	Dallas Cowboys*	10	6	0	10	6
Miami Dolphins	7	9	0	7	9	Philadelphia E+	9	7	0	9	7
Buffalo Bills	6	10	0	9	7	Washington R	7	9	0	9	7
New York Jets	4	12	0	3	13	New York G	5	11	0	6	10
AFC North						**NFC North**					
Baltimore R*	10	6	0	11	5	Chicago Bears*	12	4	0	14	2
Pittsburgh S	9	6	1	9	7	Minnesota Vikings	8	7	1	10	6
Cleveland B	7	8	1	9	7	Green Bay Packers	6	9	1	5	11
Cincinnati B	6	10	0	5	11	Detroit Lions	6	10	0	5	11
AFC South						**NFC South**					
Houston T*	11	5	0	13	3	New Orleans S*	13	3	0	9	7
Indianapolis C+	10	6	0	7	9	Carolina Panthers	7	9	0	9	7
Tennessee T.	9	7	0	8	8	Atlanta Falcons	7	9	0	8	8
Jacksonville J	5	11	0	7	9	Tampa Bay B	5	11	0	4	12
AFC West						**NFC West**					
Kansas City C*	4	0	9	7	4	Los Angeles R.*	13	3	0	12	4
Los Angeles C+	4	0	9	7	4	Seattle Seahawks+	10	6	0	11	5
Denver Broncos	10	0	8	8	10	S.Francisco 49ers	4	12	0	4	12
Oakland Raiders	12	0	4	12	12	Arizona Cardinals	3	13	0	3	13

Table 3. Model Validation

Conclusions

Major professional sports teams have been employing various Business Analytics applications and tools as they can provide competitive advantages to decision makers and teams. Team managers can employ the use of data to provide insights, make data driven decisions, and measure performance. As demonstrated in this study, predictive models such as decision trees and logistic regression analysis may be used to determine with measurable accuracy how different team statistics interact with the outcome of a game.

REFERENCES

Steinberg, Leigh. 2015. Changing the Game: The Rise of Sports Analytics, Retrieved May 2. 2019 from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#4a82a5ca4c1f.

McManus, T. (2017). Mathletes: Eagle's Analytics Team has an in Game-line to Doug Pederson, Retrieved April 4, 2019 from http://www.espn.com/blog/philadelphia-eagles/post/_/id/22272/math-movement-eagles-analytics-team-has-direct-line-to-doug-pederson-in-game.

Barnwell, B. (2017). The NFL Stats that Matter Most, Retrieved March 15, 2019 from http://www.espn.com/nfl/story/id/20114211/the-nfl-stats-matter-most-2017-offseason-bill-barnwell

Rudy, K. (2014). A Statistical Look at How Turnovers Impacted the NFL Season, Retrieved April 25, 2019, from http://blog.minitab.com/blog/the-statistics-game/a-statistical-look-at-how-turnovers-impacted-the-nfl-season.

Feng, E. (n.d.) How Passing and Rushing Affect Winning in the NFL, Retrieved April 14, 2019 from https://thepowerrank.com/2014/01/10/which-nfl-teams-make-and-win-in-the-playoffs/

Johnson, A. W. (2013). Going for Three: Predicting the Likelihood of Field Goal Success with Logistic Regression. Retrieved April 10, 2019 from https://sites.google.com/a/umich.edu/aaronwj/nfl-analytics.

JMP. (2017), Are you Ready for Some Football…Predictive Models? Retrieved April 3, 2019, from https://community.jmp.com/t5/JMP-Blog/Are-you-ready-for-some-football-predictive-models/ba-p/43546.